CHAPTER I
INTRODUCTION

In this chapter, the research background will be explained in detail, including the problems encountered and the reasons why the proposed solution is so important. Next, the main research problem and the objectives to be achieved will be discussed. Finally, the benefits of this research both in academic and industrial contexts as well as its limitations will be explained.

1.1 Background

Blood supply chain (BSC) management is one of the fundamental pillars of the national health system, which faces ongoing challenges. In Indonesia, the Indonesian Red Cross (PMI) plays a central role in managing blood, from collection to distribution. The primary challenge in BSC management is maintaining a dynamic balance between a stochastic supply and unpredictable demand [1]. Additionally, blood has a very limited shelf life, and its supply relies entirely on voluntary participation from donors, making forecasting accuracy a crucial element in daily operations [2].

The vulnerability of the blood supply chain became starkly evident during the COVID-19 pandemic, which caused extreme uncertainty in both supply and demand. This global health crisis drastically altered the landscape of blood donation worldwide, with reports indicating a decline in donation volumes of up to 20–30% across various countries [3]. This situation underscores that the ability to accurately predict blood supply is key to inventory optimization.

To address complexities that traditional forecasting methods cannot handle, modern research has turned to machine learning approaches [3]. This approach focuses on the analysis of historical and demographic data. Historical data refers to past donation trends and patterns, such as aggregated data from the previous month. Demographic data includes specific donor attributes, such as blood type, gender, and age group. A number of previous studies have confirmed that historical variables such as donation intervals and frequency, as well as demographic factors like age and gender, are strong predictors for identifying and forecasting blood donors [4] [5].

The data period for this study spans from 2019 to 2024, directly encompassing the peak phase of disruption caused by the pandemic, necessitating the development of more adaptive models. The volatility during this period renders predictive models developed in the pre-pandemic era less relevant and potentially inaccurate [3]. Case studies at various regional PMI branches in Indonesia indicate that simple methods such as Linear Regression have been applied to address this issue, as was done at the PMI Blood Donation Unit (UDD) in Bojonegoro Regency, where the model achieved an average prediction accuracy of 80.14% [2]. Similar research at the Langkat Regency PMI, which also used Linear Regression, underscores the urgent need for an effective and reliable prediction system at the local level [6].

As technology advances, machine learning offers a solution to address uncertainty in the BSC by enabling a shift from reactive strategies to a proactive, data-driven approach [1]. The application of machine learning is not limited to demand forecasting; it has also proven highly effective in improving the efficiency of outreach programs. One study successfully developed a Random Forest Classifier model to predict donor retention, achieving a Matthews Correlation Coefficient (MCC) of 0.851. This performance demonstrates a high level of accuracy in identifying donors most likely to return to donate, thereby enabling PMI to design more cost-effective campaigns [4].

Within the spectrum of machine learning algorithms, Extreme Gradient Boosting (XGBoost) has demonstrated exceptional performance and is considered a more efficient implementation of standard Gradient Boosting [7]. Various forecasting case studies have proven XGBoost's effectiveness in providing accurate predictions [8]. A study using XGBoost significantly outperformed traditional regression models, achieving a Root Mean Squared Scaled Error (RMSSE) of 0.655 a 16.3% improvement over Linear Regression [9]. Even when used in hybrid ensemble models such as Random Forest-XGBoost, it drastically improves predictive performance, with a reduction in Mean Absolute Percentage Error (MAPE) of up to 12% and an increase in the coefficient of determination (R^2) of 24% compared to previous methods [10].

XGBoost's superiority is also confirmed in various comparative studies across other domains. One study showed that XGBoost achieved 97% accuracy in regression prediction, far surpassing Linear Regression, which only reached 85% [11]. Additionally, XGBoost demonstrated perfect performance by achieving a score of 1.0 for all evaluation metrics, compared to Random Forest, which showed lower performance [12]. XGBoost's reliability is also proven in handling data with imbalanced classes and generally outperforms other Gradient Boosting methods such as LightGBM in terms of accuracy [13] [14].

To build a reliable predictive model, input feature selection is a crucial step [15]. This study develops a machine learning model to predict daily and monthly blood donor counts, classified into nine demographic targets including blood type, gender, and age group. This approach employs a series of independently trained XGBoost models utilizing historical features such as the previous month's aggregated data as well as exogenous variables relevant to the prediction day. This methodology aligns with previous findings that historical variables such as donation frequency and age are strong predictors [5]. Model evaluation is conducted at two levels daily accuracy to assess direct predictive performance and monthly aggregate accuracy to measure the model's reliability for strategic planning.

The research context is specifically set at the PMI Blood Donation Unit (UDD) in Bojonegoro Regency. This location faces challenges in meeting blood needs due to unpredictable demand, necessitating a method to ensure its availability. The case study in Bojonegoro is relevant because this district exhibits donor interest patterns that tend to be inversely related to those in major cities [2]. Previous studies have highlighted research gaps in the same location, such as the study by Sumari et al. (2021), which focused on predicting blood demand using Linear Regression. A similar approach was also applied by Putri et al. (2024) to predict the number of donors at the PMI in Langkat Regency, indicating that simple linear models remain the common approach in the context of regional PMIs in Indonesia. This research gap becomes evident when these methods are compared to more advanced approaches at the international level, which often focus on demand forecasting or use different datasets. Thus, this study focuses on predicting the number of donors by applying the XGBoost algorithm, which has proven to

outperform older methods such as Linear Regression or ARIMA in terms of blood donor prediction [3] [5] [16].

1.2 Problem Statement

Based on the background described above, the research problem in this study can be formulated as follows:

1. How does the Extreme Gradient Boosting (XGBoost) model perform in predicting the number of daily blood donors at the PMI Blood Donation Unit in Bojonegoro Regency based on blood type, gender, and age group?
2. How does the Extreme Gradient Boosting (XGBoost) model perform in predicting the monthly (aggregated) number of blood donors at the PMI Blood Donation Unit in Bojonegoro Regency based on blood type, gender, and age group?
3. How does the evaluation of predictions for the number of daily and monthly (aggregated) blood donors at the PMI Blood Donation Unit in Bojonegoro Regency, based on blood type, gender, and age group, compare?

1.3 Research Objective

The primary objective of this study is to develop and evaluate a predictive model for the daily number of blood donors at the PMI Blood Donation Unit in Bojonegoro Regency using the Extreme Gradient Boosting (XGBoost) algorithm. This study developed three separate demographic models a blood type-based model, a gender-based model, and an age group-based model, each of which was trained using a set of historical features relevant to its respective category. Furthermore, this study analyzes the performance of each model at two different levels of granularity the daily prediction level and the monthly aggregate level. The ultimate objective of this study is to evaluate and compare the relative performance of these three models to identify the most accurate and reliable modeling architecture, using the evaluation metrics Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

1.4 Benefits of Research

This study is expected to provide significant benefits to the various parties involved, including:

1. For the Author

   The benefits of this research for the author include the opportunity to apply the knowledge gained, particularly in the fields of data science and machine learning, as well as to gain deeper insights into blood supply chain management.

2. For the Institution

   The results of this research can serve as a decision-making tool to plan donor recruitment strategies in a more proactive and efficient manner. This model is expected to provide insights for optimizing blood inventory management, thereby reducing the risk of both shortages and surpluses.

3. For Future Researchers

   This study can serve as a reference and foundation for future research, particularly regarding the application of more complex machine learning algorithms in the healthcare management domain. Additionally, this study provides empirical evidence regarding the added value of using demographic data, which can be further explored.

4. For Technology Practitioners

   For technology practitioners, this study can serve as a reference for developing data driven applications for the social and health sectors. For policymakers, the findings of this study can support the formulation of more modern and effective data-driven blood supply management strategies.

5. For Technology Practitioners

   For technology practitioners, this study can serve as a reference for developing data-driven applications for the social and health sectors. For policymakers, the findings of this study can support the formulation of more modern and effective data-driven blood supply management strategies.

1.5 Problem Scope

To ensure that this study can be conducted in a focused and measurable manner, the following research boundaries have been established:

1. The study was conducted at the Indonesian Red Cross (PMI) Blood Donor Unit in Bojonegoro Regency.
2. The data used consists of historical and demographic data on 146,447 donors, covering the period from January 2019 to December 2024.
3. The demographic features used in the modeling were limited to blood type, gender, and age of the donors. It should be noted that the blood type classification did not take into account the Rh factor.